MAT 115 Midterm Project

For this midterm project you are going to practice all of the major components of a the skills required to be a data scientist. You will find a dataset that addresses a question you are interested in exploring, wrangle the dataset appropriately, visualize interesting patterns that address the question, and present it in an interesting and accessible manner.

As per usual, for this assignment you should write your answers in RMarkdown but submit both Markdown and knit file. The due date for this assignment is **Thursday, Oct. 21 at 11:59 PM**.

Data:

1) Find a dataset that interests you! If you care, then this project will be more fun for you to complete and for me to grade.

2) Data can come from many sources:

- R package (there are more packages with data than any one person could ever want)
- Kaggle (link) (data can be wrangled there by anyone but it must be attributed to a reputable source)
- Scientific data repositories like Dryad (link)

3) Dataset should be large and complex. It should include at **LEAST 5 variables and ~1,000 rows**. Exceptions can be made, but you must make a compelling argument and get pre-approval from me.

4) Data cannot be fully wrangled for you. Your question must require *meaningful* manipulation of the data that includes something like indexing, ordering, subsetting, mutating, etc. Your RMarkdown file should include all code used to manipulate the data.

5) Try to use data that is from a reputable source. You know the old saying: garbage in, garbage out. Let's avoid the issue of "junk" science. If you are unsure about a dataset, check with me.

6) I prefer the data to be real and NOT fabricated. Kaggle has a lot of fabricated datasets. Please do *NOT* use them. Again, never say never, so if you have a compelling reason to use a fabricated dataset what is the harm is asking!

7) You *must* upload the raw data to Canvas as a separate file with your report. If it is an exceptionally large dataset you can upload a small representative sample (~500 rows).

Writing:

1) What is your question and why should we care (ie, motivation)?

2) A brief description of what you did with the data. How did you wrangle it? Why did you pick the type of visualizations that you did?

3) Clear and compelling interpretation of the data that sheds *meaningful* light on your question. This should include **AT LEAST two visualizations** (with associated code) that are connected to each other in some way. One plot expands upon the other, addresses the same variables in a different way, etc.

- You should also include **2 citations** that support your interpretation of the results. These do not have to be peer-reviewed scientific papers (though those would be great!). Other acceptable sources are ".gov" or ".org" sites, along with data driven journal articles like those found on a website such as 538 (link). I do *not* care what format you use for your citations.

4) Address any potential ethics issues. Look back at our previous explorations of ethics to review the parameters of this topic.

Note: All 4 of these written components can be completed in 1 or 2 paragraphs, each.

Presentation:

You will NOT be presenting these results in class. However, you will for the final project of the semester and I want you to start exploring this skill.

There are many different forms (eg, Microsoft PowerPoint) of presentations that can be produced using RMarkdown. Here (link) are some really good resources on how to produce them. While you CAN produce the presentation in RMarkdown, you are **not require to use RMarkdown** to produce your presentation.

For this assignment, I want you to produce a 3-4 slide presentation that focuses on the material in questions #1 and #3 above in the "**Writing**" section of this assignment prompt. Make it short, simple, and compelling - include no code. Submit this as a separate document to the Canvas assignment. Use the PowerPoint presentations I have given as examples.

You are allowed to work as an individual or with a partner on this assignment. I really like the opportunity collaboration gives to improving projects like this. However, I know that many people prefer to work alone, so I am giving you the choice.

If you work with a partner, I am going to require that you join a group with your partner on Canvas (BEFORE you turn anything in) so that you are both linked to the documents that you submit. I am also requiring you fill out a group evaluation form (Canvas link). You should turn this in as a separate document to the Canvas assignment.